

PRACTICE QUESTIONS

CertiProf

CPEFPC

Prompt Engineering Foundation

SHARED BY [u/Happy-Robin-7607](#) on aiexamcerts

EDITION	Free Set 2
QUESTIONS	10
SHARED	October 15, 2026
EXAM	CertiProf CPEFPC, Prompt Engineering Foundation

Shared by [u/Happy-Robin-7607](#) on aiexamcerts

10 practice questions for the CertiProf CPEFPC exam, Prompt Engineering Foundation, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/certiprof/cpefpc. Spotted a mistake? Email support@aiexamcerts.com with the exam code CPEFPC and the question number.

1. Why can two identical prompts produce slightly different answers?

- A The model keeps a separate memory for each user and quietly changes its answer over time
- B The server clock changes the model's weights while the request is being processed by the cluster
- C Sampling randomness, such as a temperature above zero, makes each generation vary
- D The browser silently rewrites the prompt text before it is sent to the model API

Answer: C

When the model samples from its probability distribution, a temperature above zero lets different tokens be chosen on different runs, so the output varies. Keeping a separate memory per user is a tempting explanation, but it does not change the random sampling. The server clock and the browser do not alter the weights or the prompt text.

2. A standard LLM updates its trained weights each time a user tells it a new fact in a chat.

- A True
- B False

Answer: B (False)

A standard model does not change its trained weights during a normal chat, so a fact supplied in the conversation only applies while it stays in context. Some products add a separate memory feature, which is a tempting confusion with weight updates. Fine-tuning is a separate, deliberate training process.

3. Why use delimiters, such as triple quotes or XML-style tags, around user-supplied text in a prompt?

- A To make the text render in bold in the chat window for the reader's convenience
- B To compress the text to fewer tokens before it is sent to the model API
- C To separate instructions from content so the model can tell them apart
- D To encrypt the content before it reaches the model so that it stays private

Answer: C

Delimiters mark where the user-supplied content starts and ends, which helps the model separate it from the instructions around it. Bold formatting is a display effect, and it does not change how the model reads the text. Compression and encryption are different techniques that delimiters do not provide.

4. Putting a role such as 'You are a patient tutor' at the start of a prompt is called role _____.

Answer: prompting

Role prompting assigns the model a persona or expertise to shape its tone and focus. Zero-shot prompting means giving no examples, which is a tempting but different label. The role is set in the prompt rather than in the model's weights.

5. Which two settings most directly affect the length or the variety of a reply? (Choose two.)

- A The model's version number
- B Temperature
- C The language of the user's browser
- D The number of examples in the prompt
- E Max tokens

Answer: B, E

Max tokens limits the length of a reply and temperature controls the variety of token choices. The model version, browser language and example count do not set reply length or randomness directly, even though they can look relevant.

6. Which prompting method elicits step-by-step intermediate reasoning before the model states its final answer?

- A Prompt chaining without any reasoning
- B Stop sequencing
- C Chain-of-thought prompting
- D Embedding search

Answer: C

Chain-of-thought prompting asks the model to work through intermediate steps before it states the final answer, which often helps with multi-step problems. Prompt chaining splits a task across several prompts, which is a tempting neighbor but not the same. Stop sequencing and embedding search are unrelated to reasoning steps.

7. Which technique samples several independent reasoning paths and keeps the most common final answer?

- A Zero-shot prompting
- B Role prompting
- C Retrieval augmentation
- D Self-consistency

Answer: D

Self-consistency generates several independent reasoning paths and selects the answer that appears most often. Zero-shot prompting uses no examples, which is a tempting label but describes a different technique. Role prompting and retrieval augmentation do not vote across reasoning paths.

8. A temperature of 0 guarantees identical output on every run in every system.

- A True
- B False

Answer: B (False)

A temperature of 0 makes sampling close to greedy, which reduces variation, but it does not guarantee identical output across all systems and versions. Hardware, batching and model updates can still change results. Teams should not rely on temperature alone for exact reproducibility.

9. Which practice helps keep an important instruction effective in a long prompt?

- A Placing it in the middle of a long document only, between paragraphs of other text
- B Writing it in a footnote at the bottom of the input, after all the other content
- C Repeating or restating key constraints near the end of the prompt
- D Removing all punctuation from the instruction so that it reads as one long phrase

Answer: C

Restating key constraints near the end keeps them close to the point where the model writes its answer. Burying the instruction in the middle of a long document is a tempting but weaker placement. Footnotes and unpunctuated text make the instruction harder to follow.

10. What is a common reason to summarize earlier conversation turns?

- A To increase the number of tokens billed for each request made by the user account
- B To change the model's temperature automatically as the conversation grows longer
- C To delete the user's account history permanently from the platform's database
- D To free context window space while keeping the essential facts

Answer: D

Summaries free space in the context window while keeping the facts the conversation still needs. Billing more tokens is a tempting but wrong motive, since summaries usually reduce the tokens sent. Changing temperature and deleting history are unrelated actions.